Understanding Failures in Out-of-Distribution Detection with Deep Generative Models

Year: 2021 | Citations: 116 | Authors: Lily H. Zhang, Mark Goldstein, R. Ranganath

---

Abstract

Deep generative models (dgms) seem a natural fit for detecting out-of-distribution (ood) inputs, but such models have been shown to assign higher probabilities or densities to ood images than images from the training distribution. In this work, we explain why this behavior should be attributed to model misestimation. We first prove that no method can guarantee performance beyond random chance without assumptions on which out-distributions are relevant. We then interrogate the typical set hypothesis, the claim that relevant out-distributions can lie in high likelihood regions of the data distribution, and that ood detection should be defined based on the data distribution's typical set. We highlight the consequences implied by assuming support overlap between in- and out-distributions, as well as the arbitrariness of the typical set for ood detection. Our results suggest that estimation error is a more plausible explanation than the misalignment between likelihood-based ood detection and out-distributions of interest, and we illustrate how even minimal estimation error can lead to ood detection failures, yielding implications for future work in deep generative modeling and ood detection.